DOCUMENT RESUME

ED 078 023 TM 002 838

TITLE	Reliability and Confidence.
INSTITUTION	Psychological Corp., New York, N.Y.
PUB DATE	May 52
NOTE	6p.; Reprint
AVAILABLE FROM	Test Service Bulletin, The Psychological Corporation, New York, N.Y.
JOURNAL CIT	Test Service Bulletin; n44 p2-7 May 1952
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Bulletins; Correlation; Scores; Statistical Analysis; *Test Interpretation; *Test Reliability; Test Selection

ABSTRACT

Some aspects of test reliability are discussed. Topics covered are: (1) how high should a reliability coefficient be?; (2) two factors affecting the interpretation of reliability coefficients--range of talent and interval between testings; (3) some common misconceptions--reliability of speed tests, part vs. total reliability, reliability for what group?, and test reliability vs. scorer reliability; and (4) a practical checklist. (For related documents, see TM 002 839-840.) (KM)

Test Service Bulletin

Nos. 44-46 THE PSYCHOLOGICAL CORPORATION 1952-1954
GEORGE K. BENNETT, *President*

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of interest. Address communications to 304 East 45th Street, New York 17, N. Y.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

JEROME E. DOPPELT
Assistant Director

DOROTHY M. CLENDENEN
Assistant Director

ALEXANDER G. WESMAN
Associate Director of the Test Division

JAMES H. RICKS, JR.
Assistant Director

ESTHER R. HOLLIS
Advisory Service

Reprints of articles from earlier issues

44. Reliability and Confidence **2**

45. Better Than Chance **8**

46. The Correction for Guessing **13**

Test Service Bulletin

No. 44 THE PSYCHOLOGICAL CORPORATION May, 1952

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of the Test Division. Address communications to 522 Fifth Avenue, New York 36.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

ALEXANDER G. WESMAN
Associate Director of the Test Division

JEROME E. DOPPELT
MARJORIE GELINK
JAMES H. RICKS, JR.
Assistant Directors

RELIABILITY AND CONFIDENCE

THE chief purpose of testing is to permit us to arrive at judgments concerning the people being tested. If those judgments are to have any real merit, they must be based on dependable scores – which, in turn, must be earned on dependable tests. If our measuring instrument is unreliable, any judgments based on it are necessarily of doubtful worth. No one would consider relying on a thermometer which gave readings varying from 96° to 104° for persons known to have normal temperatures. Nor would any of us place confidence in measurements of length based on an elastic ruler. While few tests are capable of yielding scores which are as dependable as careful measurements of length obtained by use of a well-marked (and rigid!) ruler, we seek in tests some satisfactory amount of dependability – of "rely-ability."

It is a statistical and logical fact that no test can be valid unless it is reliable; knowing the reliability of a test in a particular situation, we know the limits beyond which validity in that situation cannot rise. Knowing reliability, we know also how large a band of error surrounds a test score – how precisely or loosely that score can be interpreted. In view of the importance of the concept of reliability, it is unfortunate that so many inadequacies in the reporting and use of reliability coefficients are to be found in the literature. This article is intended to clarify some aspects of this very fundamental characteristic of tests.

Reliability coefficients are designed to provide estimates of the consistency or precision of measurements. When used with psychological tests, the coefficients may serve one or both of two purposes: (1) to estimate the precision of the test itself as a measuring instrument, or (2) to estimate the consistency of the examinees' performances on the test. The second kind of reliability obviously embraces the first. We can have unreliable behavior by the examinee on a relatively reliable test, but we cannot have reliable performance on an unreliable instrument. A student or applicant suffering a severe headache may give an uncharacteristic performance on a well-built test; the test may be reliable, but the subject's performance is not typical of him. If, however, the test items are ambiguous, the directions are unclear, or the pictures are so poorly reproduced as to be unintelligible – if, in short, the test materials are themselves inadequate – the subject is prevented from performing reliably, however propitious his mental and physical condition.

This two-fold purpose of reliability coefficients is reflected in the several methods which have been developed for estimating reliability. Methods which provide estimates based on a single sitting offer evidence as to the precision of the test itself; these include internal consistency estimates, such as those obtained by use of the split-half and Kuder-Richardson techniques when the test is given only once, as well as estimates based on immediate retesting, whether with the same form or an equivalent one. When a time interval of one or more days is introduced, so that day-to-day variability in the person taking the test is allowed to have an effect, we have evidence concerning the stability of the trait and of the examinee as well as of the test. It is important to recognize whether a reliability coefficient describes only the test, or whether it describes the stability of the examinees' performances as well.

THE PSYCHOLOGICAL CORPORATION believes that tests should be bought on the basis of their *quality as measuring instruments* and their *appropriateness for the user's purpose.*

This article is one of a series offered to help counselors, personnel men, psychologists, psychiatrists, and educators to a fuller understanding of mental measurements so that they can choose tests more wisely and use them more effectively. Previous issues of this *Bulletin* include:

No. 36 on the concept of aptitudes
No. 37 on the validation of tests
No. 38 on the use of expectancy tables
No. 39 on norms
No. 40 on correlation coefficients
Nos. 41 and 43 on the identification of children's special abilities
No. 42 on the cost of testing

Upon request, THE PSYCHOLOGICAL CORPORATION will be glad to send copies of any of these earlier *Bulletins* without charge.

How High Should a Reliability Coefficient Be?

We should naturally like to have as much consistency in our measuring instruments as the physicist and the chemist achieve. However, the complexities of human personality and other practical considerations often place limits on the accuracy with which we measure and we accept reliability coefficients of different sizes depending on various purposes and situations. Perhaps the most important of these considerations is the gravity of the decision to be made on the basis of the test score. The psychologist who has to recommend whether or not a person is to be committed to an institution is obligated to seek the most reliable instruments he can obtain. The counselor inquiring as to whether a student is likely to do better in one curriculum or another may settle for a slightly less reliable instrument, but his demands should still be high. A survey of parents' attitudes towards school practices needs only moderate reliability, since only the *average* or group figures need to be highly dependable and not the specific responses of individual parents. Test constructors experimenting with ideas for tests may accept rather low reliability in the early stages of experimentation – those tests which show promise can then be built up into more reliable instruments before publication.

It is much like the question of how confident we wish to be about decisions in other areas of living. The industrial organization about to hire a top executive (whose decisions may seriously affect the entire business) will usually spend large sums of time and money to obtain reliable evidence concerning a candidate's qualifications for the job. The same firm will devote far less time or money to the hiring of a clerk or office boy, whose errors are of lesser consequence. In buying a house, we want to have as much confidence in our decision as we can reasonably get. In buying a package of razor blades, slim evidence is sufficient since we lose little if we have to throw away the entire package or replace it sooner than expected. The principle is simply stated: the more important the decision to be reached, the greater is our need for confidence in the precision of the test and the higher is the required reliability coefficient.

Two Factors Affecting the Interpretation of Reliability Coefficients

Actually, there is no such thing as *the* reliability coefficient for a test. Like validity, reliability is specific to the group on which it is estimated. The reliability coefficient will be higher in one situation than in another according to circumstances which may or may not reflect real differences in the precision of measurement. Among these factors are the range of ability in the group and the interval of time between testings.

Range of talent

If a reliability estimate is based on a group which has a small spread in the ability measured by the test, the coefficient will be relatively low: If the group is one which has a wide range in that particular talent, the coefficient will be higher. That is, the reliability coefficient will vary with the range of talent in the group, even though the accuracy of measurement is unchanged. The following example may illustrate how this comes about. For simplicity, we have used small numbers of cases; ordinarily, far larger groups would be required to ensure a coefficient in which we could have confidence.

In Table I are shown the raw scores and rankings of twenty students on two forms of an arithmetic test. Looking at the two sets of rankings, we see that changes in rank from one form to the other are minor; the ranks shift a little, but not importantly. A coefficient computed from these data would be fairly high.

TABLE I. Raw Scores and Ranks of Students on Two Forms of an Arithmetic Test

Student	Form X		Form Y	
	Score	Rank	Score	Rank
A	90	1	88	2
B	87	2	89	1
C	83	3	76	5
D	78	4	77	4
E	72	5	80	3
F	70	6	65	7
G	68	7	64	8
H	65	8	67	6
I	60	9	53	10
J	54	10	57	9
K	51	11	49	11
L	47	12	45	14
M	46	13	48	12
N	43	14	47	13
O	39	15	44	15
P	38	16	42	16
Q	32	17	39	17
R	30	18	34	20
S	29	19	37	18
T	25	20	36	19

Now, however, let us examine only the rankings of the five top students. Though for these five students the shifts in rank are the same as before, the importance of the shifts is greatly emphasized. Whereas in the larger group student C's change in rank from third to fifth represented only a ten per cent shift (two places out of twenty), his shift of two places in rank in the smaller top group is a forty per cent change (two places out of five). When the entire twenty represent the group on which we estimate the reliability of the arithmetic test, going from third on form X to fifth on form Y still leaves the student as one of the best in this population. If, on the other hand, reliability is being estimated only on the group consisting of the top five students, going from third to fifth means dropping from the middle to the bottom of this population — a radical change. A coefficient, if computed for just these five cases, would be quite low.

Note that it is not the smaller number of cases which brings about the lower coefficient. It is the narrower range of talent which is responsible. A coefficient based on five cases as widespread as the twenty (e.g., pupils A, E, J, O, and T, who rank first, fifth, tenth, fifteenth, and twentieth respectively on form X), would be at least as large as the coefficient based on all twenty students.

This example shows why the reliability coefficient may vary even though the test questions and the stability of the students' performances are unchanged. A test may discriminate with satisfactory precision among students with wide ranges of talent but not discriminate equally well in a narrow range of talent. A yardstick is unsatisfactory if we must differentiate objects varying in length from 35.994 to 36.008 inches. Reliability coefficients reflect this fact, which holds regardless of the kind of reliability coefficient computed. It should be obvious, then, that *no reliability coefficient can be properly interpreted without information as to the spread of ability in the group on which it is based.* A reliability coefficient of .65 based on a narrow range of talent is fully as good as a coefficient of .90 based on a group with twice that spread of scores. Reliability coefficients are very much a function of the range of talent in the group.

Interval between testings

When two forms of a test are taken at a single sitting, the reliability coefficient computed by correlating the two forms is likely to overestimate somewhat the real accuracy of the test. This is so because factors such as mental set, physical condition of examinees, conditions of test administration, etc. — factors which are irrelevant to the test itself — are likely to operate equally on both forms, thus making each person's pair of scores more similar than they otherwise would be. The same type of overestimate may be expected when reliability is computed by split-half or other internal consistency techniques, which are based on a single test administration. Coefficients such as these describe the accuracy of the test, but exaggerate the practical accuracy of the results by the extent to which the examinees and the testing situation may normally be expected to fluctuate. As indicated above, coefficients based on a single sitting do not describe the stability of the subjects' performances.

When we set out to investigate how stable the test results are likely to be from day to day or week to week, we are likely to underestimate the test's accuracy, though we may succeed in obtaining a realistic estimate of stability of the examinees' performances on the test. The underestimation of the test's accuracy depends on the extent to which changes in the examinees have taken place between testings. The same influences mentioned above — mental set, physical condition of examinees, and the like — which *increase* coefficients based on a single sitting are likely to *decrease* coefficients when testing is done on different days. It is unlikely for example, that the same persons who had headaches the first day will also have headaches on the day of the second testing.

Changes in the persons tested may also be of a kind directly related to the content of the particular test. If a month has elapsed between two administrations of an arithmetic test, different pupils may have learned different amounts of arithmetic during the interval. The second testing should then show greater score increases for those who learned more than for those who learned less. The correlation coefficient under these conditions will reflect the test's accuracy *minus* the effect of differential learning; it will not really be a reliability coefficient.

For most educational and industrial purposes, the reliability coefficient which reflects *stability of performance* over a relatively short time is the more important. Usually, we wish to know whether the student or job applicant would have achieved a similar score if he had been tested on some other day, or whether he might have shown up quite differently. It would be unfortunate and unfair to make important decisions on the basis of test results which might have been quite different had the person been tested the day before or a day later. We want an estimate of reliability which takes into account

accidental changes in day-to-day ability of the individual, but which has not been affected by real learning between testings. Such a reliability coefficient would be based on two sittings, separated by one or more days so that day-to-day changes are reflected in the scores, but not separated by so much time that permanent changes, or learning, have occurred.* Two forms of a test, administered a day to a week apart, would usually satisfy these conditions. If the same form of a test is used in both sittings, the intervening time should be long enough to minimize the role of memory from the first to the second administrations.

Ideally, then, our reliability coefficient would ordinarily be based on two different but equivalent forms of the test, administered to a group on two separate occasions. However, it is often not feasible to meet these conditions: there may be only one form of the test available, or the group may be available for only one day, or the test may be one which is itself a learning experience. We are then forced to rely on coefficients based on a single administration. Fortunately, when such coefficients are properly used they usually provide close approximations to the estimates which would have been obtained with alternate forms administered at different times.

Some Common Misconceptions

Reliability of speed tests

Although estimates of reliability based on one administration of the test are often satisfactory, there are some circumstances in which *only* retest methods are proper. Most notable is the case in which we are dealing with an easy test given under speed conditions. If the test is composed of items which almost anyone can answer correctly given enough time but which most people tested cannot finish in the time allowed, the test is largely a measure of speed. Many clerical and simple arithmetic tests used with adults are examples of speed tests. Internal consistency methods, whether they are of the Kuder-Richardson or of the split-half type, provide false and often grossly exaggerated estimates of the reliability of such tests. To demonstrate this problem, two forms of a simple but speed-laden clerical test were given to a group. For *each* form the odd-even (split-half) reliability coefficient was found to be over .99. However, when scores on Form A were correlated with scores on Form B, the coefficient was .88. This latter value is a more accurate estimate of the reliability of the test.* Many equally dramatic illustrations of how spurious an inappropriate coefficient can be may be found readily, even in manuals for professionally made tests.

If a test is somewhat dependent on speed, but the items range in difficulty from easy to hard, internal consistency estimates will not be as seriously misleading as when the test items are simple and the test is highly speeded. As the importance of speed diminishes, these estimates will be less different from the coefficients which would be obtained by retest methods. It is difficult to guess how far wrong an inappropriate coefficient for a speeded test is. ***Whenever there is evidence that speed is important in test performance, the safest course is to insist on an estimate of reliability based on test-and-retest,*** if necessary with the same but preferably with an alternate form of the test.

Part vs. total reliability

Some of the tests we use are composed of several parts which are individually scored and the part scores are then added to yield a total score. Often, reliability is reported only for the total score, with no information given as to the reliability of the scores on the individual parts. This may lead to seriously mistaken assumptions regarding the reliability of the part scores – and, thus, of the confidence we may place in judgments based on the part scores. The longer a test is, other things being equal, the more reliable it is; the shorter the test, the lower is its reliability likely to be. A part score based on only a portion of the items in a test can hardly be expected to be as reliable as the total score; if we treat the part score as though it has the reliability of the total score, we misplace our confidence – sometimes quite seriously.

As an example, we may look at the Wechsler Intelligence Scale for Children, one of the most important instruments of its kind. Five subtests are combined to yield a total Verbal Score for this test. The reliability coefficient for the Verbal Score, based on 200 representative ten-year-olds, is .96 – high enough to warrant considerable confidence in the accuracy of measurement for these youngsters. For the same population, however, a single subtest (General Comprehension) yields a reliability coefficient of only .73 – a far less impressive figure. If we allow our-

* A coefficient which is based on two testings between which opportunity for learning has occurred is a useful statistic. It may provide evidence of how much individual variation in learning has taken place, or of the stability of the knowledge, skills or aptitudes being measured. It is similar to a reliability coefficient, and is in part a function of the reliability of the two measurements; but such a coefficient should not be interpreted as simply estimating reliability – it requires a more complex interpretation.

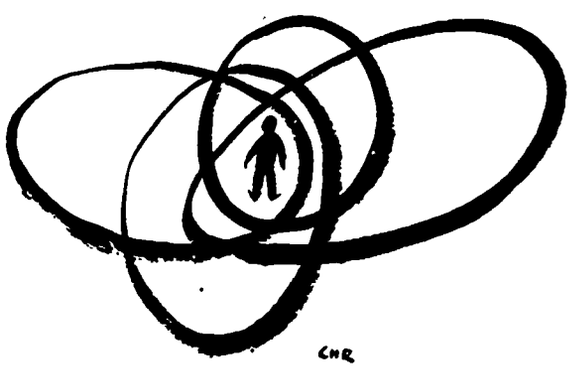
* Manual for the *Differential Aptitude Tests*, Revised Edition, page 65. The Psychological Corporation, 1952.

selves to act as though the total test reliability coefficient of .96 represents the consistency of measurement we can expect from the Comprehension subtest, we are likely to encounter unpleasant surprises on future retests. More importantly, any clinical judgments which ignore the relatively poor reliability of the part score are dangerous. Test users should consider it a basic rule that *if evidence of adequate reliability for part scores is missing, the part scores should not be used.*

Reliability for what group?

This question may be considered as a special case under the principles discussed above with respect to range of talent. It is worth special consideration because it is so often ignored. Even the best documented of test manuals present only limited numbers of reliability coefficients; in too many manuals a single coefficient is all that is made available. On what group should a reliability coefficient be based?

When we interpret an individual's test score, the most meaningful reliability coefficient is one based on the group with which the individual is competing. Stated otherwise, the most appropriate group is that in which the counselor, clinician or employment



Each of us is a member of many groups

manager is trying to make decisions as to the relative ability of the individuals on the trait being measured. Any one person is, of course, a member of many groups. An applicant for a job may also be classified as a high school or college graduate, an experienced or inexperienced salesman or bookkeeper, a local or out-of-state person, a member of one political party or another, below or above age thirty, etc. A high school student is a boy or girl; a member of an academic, trade or commercial school group; a member of an English class, a geometry class, or a woodworking or cooking class; a freshman or a junior; a future engineer or nurse or garage mechanic. Obviously, it would be impossible for a test manual to offer reliability for *all* the groups of which any one individual is a member.

The appropriate group is represented by the individual's present competition. If we are testing applicants for clerical work, the most meaningful reliability coefficient is one based on applicants for clerical work. Coefficients based on employed clerical workers are somewhat less useful, those based on high school graduates are still less useful; as we go on to *more general* groups – e.g., all high school students or all adults – the coefficients become less and less meaningful. Similarly, as we go to *less relevant* groups (even though they may be quite specific) the reliability coefficients are also less relevant and less meaningful. The reliability of a test calculated on the basis of mechanical apprentices, college sophomores, or junior executives reveals little of importance when we are concerned with clerical applicants. What we need to know is how well the test discriminates among applicants for clerical work. If we can define the population with even greater specificity and relevance –e.g., female applicants for filing jobs–so much the better. *The closer the resemblance between the group on which the reliability coefficient is based and the group of individuals about whose relative ability we need to decide, the more meaningful is that coefficient of reliability.*

Test reliability vs. scorer reliability

Some tests are not entirely objective as to scoring method; the scorer is required to make a judgment as to the correctness or quality of the response. This is frequently true in individually-administered tests (Wechsler or Binet for example), projective techniques in personality measurement (Rorschach, Sentence Completion, etc.) and many other tests in which the subject is asked to supply the answer, rather than to select one of several stated choices. For tests such as these, it is important to know the extent of agreement between the persons who score them. Test manuals usually report the amount of agreement by means of a coefficient of correlation between scores assigned to a set of test papers by two or more independent scorers.

Such a correlation coefficient yields important information–it tells us how objectively the test can be scored. It even contributes some evidence of reliability, since objectivity of scoring is a factor which makes for test reliability. Such a coefficient should not, however, be considered a reliability coefficient for the test; it is only an estimate of *scoring* reliability –a statement of how much confidence we may have that two scorers will arrive at similar scores for a given test paper. Moreover, it is possible for a test to be quite unreliable as a measuring instrument, yet have high scoring objectivity. We should remember that many objective tests–those in which the person

selects one of several stated options—are not very reliable, yet the scoring is by definition objective. A short personality inventory may have a retest reliability coefficient of .20; but if it is the usual paper-and-pencil set of questions with a clear scoring key, two scorers should agree perfectly, except for clerical errors, in assigning scores to the test. The coefficient of correlation between their sets of scores might well be 1.00.

In short, information as to scorer agreement is important but not sufficient. The crucial question—How precisely is the test measuring the individual?—is not answered by scorer agreement; a real reliability coefficient is required.

A Practical Check-list

When reading a test manual, the test user would do well to apply a mental check-list to the reliability section, raising at least the following questions for each reliability coefficient:

1. What does the coefficient measure?
   a. Precision of the test—coefficient based on single sitting?
   b. Stability of examinees' test performances—coefficient based on test-and-retest with a few days intervening?
2. Is it more than a reliability coefficient? . . . . does it also measure constancy of the trait? . . . . is the coefficient based on test-and-retest with enough intervening time for learning or similar changes to have occurred?
3. Do scores on the test depend largely on how rapidly the examinees can answer the questions? If so, is the reliability coefficient based on a test-and-retest study?
4. Are there part scores intended for consideration separately? If so, is each part score reliable enough to warrant my confidence?
5. Is the group on which this coefficient is based appropriate to my purpose? Does it consist of people similar to those with whom I shall be using the test?
6. Since a reliability coefficient, like any other statistic, requires a reasonable number of cases to be itself dependable, how large is the group on which the coefficient is based?

If, and *only* if, the coefficients can be accepted as meeting the above standards, one may ask:

7. In view of the importance of the judgments I shall make, is the correlation coefficient large enough to warrant my use of the test?

A reliability coefficient is a statistic—simply a number which summarizes a relationship. Before it takes on meaning, its reader must understand the logic of the study from which the coefficient was derived, the nature of the coefficient and the forces which affect it. Statistics may reveal or conceal—what they do depends to a very large extent on the logical ability and awareness the reader brings to them. Figures do lie, to those who don't or won't understand them.

—A.G.W.

A screening and counseling aid of interest to high school and college users

●SURVEY OF STUDY HABITS AND ATTITUDES

WILLIAM F. BROWN AND WAYNE H. HOLTZMAN, University of Texas

To counselors and educators, the study habits and attitudes of their students are of immense importance. It is in these terms that one most often seeks the explanation of the well-endowed student who earns only poor grades while others with mediocre scholastic aptitude are achieving a better record. The Brown-Holtzman *Survey of Study Habits and Attitudes* (*SSHA*) is designed to detect cases in which this is a likely source of difficulty in college and to help these students and their advisors plan steps which may avert the difficulty.

In taking the *SSHA*, the student indicates the frequency with which he practices or the extent to which he agrees with each of seventy-five study procedures or beliefs. The scoring keys reflect systematic development and rigorous cross-validation against actual grades earned in ten colleges, from Amherst to U.C.L.A. As a result, the *SSHA* can be used effectively as

a) a *screening device*, to identify among freshmen entering college those most likely to need early preventive help;
b) a *diagnostic instrument* and counseling aid, by use of a special Counseling Key which can be laid over the student's answer sheet to indicate specific practices or beliefs which may handicap him;
c) a *teaching aid*, not only in remedial or how-to-study classes but also in elementary psychology and education courses where it stimulates lively discussion both of scores and of the statements which make up the *Survey;* and
d) a *research tool*, in investigations of the learning or the counseling processes.

Test Service Bulletin

No. 45 THE PSYCHOLOGICAL CORPORATION May, 1953

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of the Test Division. Address communications to 522 Fifth Avenue, New York 36.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division
ALEXANDER G. WESMAN
Associate Director of the Test Division

JEROME E. DOPPELT
MARJORIE GELINK
JAMES H. RICKS, JR.
Assistant Directors

BETTER THAN CHANCE

"TESTS with a coefficient of validity less than .50 are practically useless, except in distinguishing between extreme cases, since at that value of r the forecasting efficiency is only 13.4 per cent."[1] This statement is quoted from one of the leading statistical texts; its paraphrase may be found in many other texts, in doctoral dissertations and other treatises of greater or lesser authority. But relatively few validity coefficients, especially in industry, exceed .50.

Why are tests being used even though they generally fall into this "practically useless" class? Is it because of ignorance or the part of test users? Not at all. Witness the statement by the author of the above quotation in reviewing a [illegible]ith validity coefficients averaging .35 to .55 in various institutions: "[the test] has shown substantial valu[illegible]redicting scholarship at the graduate level."[2] Now the forecasting efficiency (using the same formula as was used above) even when $r = .55$ is about 16 per cent — hardly enough to warrant the author's shift from condemnation to commendation. The reader might justifiably be confused — if the expert can't agree with himself, what is the counselor or personnel man to think?

Reassurance is in order. The test user may follow the practice of the expert, without violating the principle enunciated in the texts. The "index of forecasting efficiency" as formulated in the texts is concerned with a precision of prediction much finer than that required in most practical situations. As a measure of the real utility of a test, the index may be grossly misleading. A more crucial consideration is the extent to which broader judgments are improved.

The difference between the two concepts can be seen clearly if we consider the prediction, in two different situations, of how far several men can broadjump. If the occasion is an athletic contest, we might want to predict just how many feet and inches each man will cover. The average difference between our estimated distances and the actual jumps will serve as a crude indicator of our predictive efficiency — the better (i.e., the more valid) the basis on which we make our predictions, the smaller this average difference will become, and the percentage by which it decreases is analogous to the per cent of improvement over chance. But suppose we move from the athletic contest or theoretical laboratory situation to one in which the practical values are extreme: say, one in which it's necessary to leap across a brook. Those who fail by inches to make it will get their feet wet as will those who miss by six feet. And those

[1]References denoted by superscript numbers will be found at the end of this article, page 12.

who just clear it will be as useful on the other side as those who sail over with five feet to spare. *Now* the test of the efficiency of our predictive test lies in the confidence with which it permits us to say, "of men who score like this, nine out of ten will make it, but of those whose scores are the lowest only three out of ten will get across." Of course, the absolute dichotomy is as extreme in its way as the pinpoint precision estimate is at the other extreme. But when we are trying to guess in which general category, high, middle or low — the champions, the experts, the good, the just average, or the duffers — our candidates will fall, we're closer to the second situation than the first. Most counselors, personnel men, and clinicians have to work with these cruder approximations.

Per cent of improvement over chance, as used with the index of forecasting efficiency, refers to the narrowing of a zone of error around a predicted score. When the validity coefficient is zero, knowledge of a test score does not permit us to predict an individual's score on the criterion with any accuracy at all; the best guess we can make with respect to any individual, regardless of how he scored on such a test, is that he will be average on the criterion. The band of error (the standard error of estimate) is as large as the spread (the standard deviation) of the ratings on the criterion for the entire group. As the correlation between the test scores and the criterion ratings increases, our precision in predicting ratings of individuals on the criterion also increases and we may predict with some degree of confidence, for example, that a person who scores in the top quarter on the test will be rated in the top quarter on the criterion as well. Of course, some of our predictions will be in error: i. e., some of those whose scores are in the top quarter on the test will be rated in the second quarter on performance, a smaller number in the third quarter, and a few may even be rated in the lowest quarter. The larger the validity coefficient, the fewer misplaced persons there will be; furthermore, the smaller will be the amount of displacement. In other words, if the validity coefficient is really high, we may expect most of those who score in the top quarter on the test to be rated in the top quarter on performance as well, a very few to be rated in the second quarter, and fewer still (or perhaps even none at all) to be rated in the third or fourth quarters.

The number of persons for whom statistically calculated predictions are wrong, and the amount by which our estimates are in error are reflected in the *standard error of estimate*. When validity is perfect, the standard error of estimate is zero; when validity is zero, the standard error of estimate is at its maximum. As the validity increases, the standard error of estimate decreases. The degree to which the standard error of estimate is reduced is what is meant by the textbook statements concerning improvement over chance. In this sense, large validity coefficients are necessary; it takes an $r = .866$ to cut the standard error of estimate even to half the size of the standard deviation of the criterion ratings — a "fifty per cent improvement over chance."

What permits us to use tests effectively even though their validity coefficients are considerably lower than .866? First, there is the matter of precision. The standard error of estimate refers to the band of error around predictions of precise, specific rankings of each individual on the criterion. In most practical work, such precision is unnecessary. We do not ordinarily need to predict that John Jones will be exactly at the 85th percentile in a college class, or that Bill Smith will be 19th in a group of 25 engineering apprentices. We are far more likely to be concerned with whether Jones will survive the first year in college, or whether Smith will be one of the satisfactory apprentices. For these purposes, whether Jones is at the 75th percentile or 90th percentile is of lesser moment; we can make a quite confident prediction that he will succeed, even though there may be a fair-sized standard error of estimate applicable to the specific percentile our formula predicts.

A second factor working in our favor in the practical use of tests is that, as the opening quotation notes, predictions are most accurately made at the extremes — and it is the extremes that are of greatest interest to us. Few colleges grant large scholarships to more than 10 or 20 per cent of their students. Few colleges fail as many as half their students and few industrial firms fire as many as half of those they hire. More often, the failures are 10 per cent or 20 per cent or possibly 30 per cent — the extremes. Thus a test which does not predict with accuracy whether students will be at the 40th percentile or the 60th percentile, can still do a valuable service in predicting that very few of the high scorers will be in the 20 per cent who fail during the freshman year, or that hardly any scholarship winners will be academic failures. In industrial selection, a test of moderate validity can be efficient in quickly screening out the "clearly ineligible" from the "clearly eligible." There will remain an indifferent zone of test scores for persons in the "eligible"

range; for them, other considerations than test scores may determine whether they should be hired.

Let us look at some data. One hundred ninety-one eighth-grade boys took the *Verbal Reasoning Test* of the *Differential Aptitude Tests* (DAT) battery at the start of a term. At the end of the term, the grades they earned in a Social Studies course were obtained. Seventy-six were found to have earned grades of D or lower; they represented 40 per cent of the total class. On the basis of chance (i. e., using a test with *zero* validity), we should expect to find that 40 per cent of those at each test score level — low, medium or high — obtained grades of D or lower. The coefficient of correlation between the test scores and these grades was .61, for which the index of forecasting efficiency comes out to just 20 per cent better than chance — hardly enough to notice. Table I reveals a very different story — it shows the test to

Table I. Chance expectations and actual performances in a social studies class in relation to *DAT-Verbal Reasoning* scores.

DAT Verbal Reasoning Test Score	No. of Pupils	% expected by chance to earn D, E, or F	% actually earning D, E, or F
26-up	19	40	6
18-25	49	40	14
10-17	60	40	36
2- 9	63	40	73

be a highly efficient predictor for the school's purposes! Instead of 40 per cent of the highest-scoring pupils being found in the low grades group (as one would expect by chance), only six per cent are found there.[3]

Another example, drawn from the area of industrial testing, is shown in Table II. The *Short Employment Tests (SET)* were administered to 74 stenographers at a single level of job responsibility, and the relationships between scores on the tests and on-the-job proficiency ratings were investigated. The girls were rated as low, average or high in ability; the table shows, for each of these groups, what per cent were in each third on the *Clerical* aptitude test of the *SET* battery.

By chance alone, the per cent of upper, middle and low scorers in each of the rated groups would be the same — in this case, 33⅓%. The boldface numbers in the table would consist of nine 33's. Note how closely this expected per cent is approximated for those ranked average in proficiency, and for those in the middle third on test score; the percentages in the middle row and those in the middle column run between 28 and 36. Note also that at the extremes — the four corner numbers — the prediction picture is more promising. Among those *rated* low, there are almost three times as many people from the lowest third on the test as there are from the top third. Among those rated high, the per cent from the top third on the test is almost two and one-half times as great as the per cent from the bottom third. The personnel man would do well to be guided by these data in selecting future stenographers, even though the validity coefficient is just .38.

Table II. Per cent of stenographers in each third on *SET-Clerical* who earned various proficiency ratings.

SET-Clerical Test Score	Proficiency Rating		
	Low	Average	High
Upper Third	18	33	50
Middle Third	29	36	28
Lowest Third	53	31	22
Total Per Cent	100	100	100
No. of Stenographers	17	39	18

The data in the above examples are based on relatively small numbers of cases (which is typically true of practical test situations) and the per cents found in each category are consequently somewhat unstable. The validity coefficients based on groups of such sizes are, of course, also less stable than coefficients based on large numbers of cases. The wise test user will make several validity studies using successive groups. Having done so, he may take an average of the validity coefficients from these studies as being a more dependable estimate of the validity of the test in his situation. Formal tables are available[4] which can be used to estimate expectancies when the validity coefficient is of a given size and the per cent of successes and failures is known. Table III has been constructed from these formal tables to illustrate the usefulness of coefficients of various magnitudes.

Table III. Per cent of successful individuals in each decile on test score —

Standing on the test		when the total per cent of failures is 20%, and				when the total per cent of failures is 30%, and				when the total per cent of failures is 50%, and			
Percentile	Decile	r=.30	r=.40	r=.50	r=.60	r=.30	r=.40	r=.50	r=.60	r=.30	r=.40	r=.50	r=.60
90-99th	10	92%	95%	97%	99%	86%	91%	94%	97%	71%	78%	84%	90%
80-89th	9	89	91	94	97	81	85	89	92	63	68	73	78
70-79th	8	86	89	91	94	78	81	84	88	59	62	65	69
60-69th	7	84	86	88	91	75	77	80	83	55	57	59	61
50-59th	6	82	84	85	87	72	74	75	77	52	52	53	54
40-49th	5	80	81	82	83	70	70	70	71	48	48	47	46
30-39th	4	78	77	77	78	67	66	65	64	45	43	41	39
20-29th	3	75	73	72	71	63	61	59	56	42	38	35	31
10-19th	2	71	68	64	61	59	55	50	45	37	33	28	22
1- 9th	1	63	56	49	40	50	43	35	27	29	23	16	10

The first part of Table III is based on a failure rate of 20 per cent. It shows the per cent of individuals at different levels on the test who are successful (in marks earned, or dollar sales, or merit rating, or number of widgets assembled, or whatever we are trying to predict) when the validity coefficient is .30, .40, .50, or .60. The columns in boldface at the left show the decile rank on the test — individuals with percentile ranks of 90 to 99 are in the tenth decile or top 10 per cent, those with percentile ranks from 80 to 89 are in the next (9th) decile, etc.; the first decile includes the individuals between the first and ninth percentiles on the test — the 10 per cent who scored lowest. In the first lightface column is shown the per cent of persons in each decile who may be expected to succeed when the validity coefficient (r) is .30; the second column in lightface type presents similar expectancy information when $r = .40$, the next column is for $r = .50$, and the last column for a validity coefficient of .60.

What does this table tell us? Suppose that the failure rate among Winsocki college freshmen is about 20 per cent — that usually one out of every five students fails or goes on probation before the end of the year. A selection test is given and a correlation of .30 is found between scores on the test and success in the first year. Ninety-two per cent of those who score in the top 10 per cent of the group on the test may be expected to succeed, while only 63 per cent in the bottom decile can expect to survive the first year. If the validity coefficient is .40, ninety-five per cent in the top decile may be expected to survive; of the lowest scoring students, 56 per cent are likely to be around at the end of the year. The survival rate when $r = .60$ is almost perfect (99 per cent) for the top group; it is only 40 per cent for the lowest scorers.

The last two sections of Table III present similar information for coefficients of .30, .40, .50, and .60 when failure rates are 30 per cent and 50 per cent. The last column at the right shows, for example, that if only 50 per cent of a total group is successful, and the validity coefficient is .60, the top scoring individuals will have a survival rate of 30 per cent; of those in the bottom decile on the test, only one out of ten is likely to succeed.

It is interesting to compare the figures in the column headed $r = .50$ (when failures total 20 per cent) with the quotation with which we began. The "only 13.4 per cent" sort of statement may be (and often has been) misinterpreted as indicating that the test can tell us little. Actually, the test has changed our picture dramatically. Without it, we could say only that for every person the odds are four chances to one he'll succeed. With the test, we can sort the candidates into groups and say that some have distinctly better prospects than others. If three men score, respectively, in the tenth, the seventh and the lowest deciles, we can give odds on their success:

	Without test information	With knowledge of test scores
Man in 10th decile	4 to 1	37 to 1 (97.4%-2.6%)
Man in 7th decile	4 to 1	8 to 1 (88%-12%)
Man in 1st decile	4 to 1	1 to 1 (49%-51%)

What are the practical implications of these facts? Most apparent is the real potential utility of validity coefficients of .60, .50, .40, and even .30; the information they provide is far from useless. For the counselor, they offer increased ability to estimate his client's general chances of success in an educational or vocational pursuit. For the admissions officer in a college, better forecasts of drop-out rate, as well as more informed selection, are possible. For personnel men in industry, data such as these provide information with respect to the selection ratios which will be necessary to obtain a desired number of successful employees.

As do all other statistics, standard errors of estimate and validity coefficients require full understanding. For all of us, our errors of estimate will always be greater than we would like. The precision of our estimates will be less than perfect, and we shall aim constantly to increase that precision. At the same time, if a test will increase appreciably our ability to predict (even though broadly) performance in curricula or careers, let us use it — with caution, but also with gratitude. A blade not sharp enough for shaving can still be used to cut a knot. — A. G. W.

NOTE: While this *Bulletin* was in press, another approach to the topic was published by W. L. Jenkins, "An index of selective efficiency (S) for evaluating a selection plan." *Journal of Applied Psychology*, 1953, Vol. 37, p. 78.

REFERENCES

[1]J. P. Guilford, *Psychometric Methods* (New York: McGraw-Hill, 1936) p. 364. The index of forecasting efficiency $= 100\ (1 - \sqrt{1 - r^2})$ where r is the validity coefficient, the correlation between the predictor test and the subsequent performance rating or other criterion. When the number of cases is small, a correction term $\left(\frac{N-1}{N-2}\right)$ is inserted under the square root sign.

[2]J. P. Guilford, test 304, page 407 in Buros: *The Fourth Mental Measurements Yearbook* (Highland Park, N. J.: The Gryphon Press, 1953).

[3]For semitechnical discussions of everyday ways of demonstrating test validity, see THE PSYCHOLOGICAL CORPORATION'S *Test Service Bulletins* Nos. 37 and 38: "How Effective Are Your Tests?" and "Expectancy Tables: A Way of Interpreting Test Validity."

[4]R. W. B. Jackson and A. J. Phillips, "Prediction Efficiencies by Deciles for Various Degrees of Relationship." *Educational Research Series* No. 11, Dept. of Educational Research, Ontario College of Education, University of Toronto. Especially interested persons may find it worthwhile to see also: H. C. Taylor and J. R. Russell, "The relationship of validity coefficients to practical effectiveness of tests in selection." *Journal of Applied Psychology*, 1939, Vol. 23, pp. 565-578.

Long awaited . . . sorely needed . . . indispensable . . .

THE FOURTH MENTAL MEASUREMENTS YEARBOOK

OSCAR K. BUROS, Editor, with 308 REVIEWERS

This is the latest in the series of the Buros *Mental Measurements Yearbooks*, which have become one of the most important references in the test field. Teachers, counselors, clinicians, personnel men — all serious test users, in fact — have found these volumes unique in their wealth of evaluative information and in the exhaustive reference listings they contain. Here are expert reviews of achievement tests, aptitude tests, individual and group intelligence tests, interest inventories, and measures of character and personality; tests of reading and tests of etiquette; of Latin and Greek, and health and home economics; of hearing, of manual dexterity and of aptitude for law school. Many of the reviews are sharply critical. Appropriate alarms are sounded for the weak points of many a test and test manual.

As summarized in its prospectus, "*The Fourth Yearbook*, a large volume of 1,189 two-column pages, consists entirely of new material and supplements rather than supplants previous yearbooks. [It] covers the period 1948 through 1951. The section "Tests and Reviews" lists 793 tests, 596 reviews by 308 reviewers, 53 excerpts from test reviews in 15 journals, and 4,417 references on the construction, validity, use and limitations of specific tests. The section "Books and Reviews" lists 429 books on measurement and closely related fields, and 758 excerpts from book reviews in 121 journals."

Those who reviewed *The Third Mental Measurements Yearbook* used such adjectives as "monumental," "indispensable," "invaluable," "comprehensive." We have no doubt that *The Fourth Yearbook* will receive equal acclaim. No school, clinic or personnel office can afford to be without it. Its value to test users many times exceeds its cost.

THE THIRD MENTAL MEASUREMENTS YEARBOOK (1948), covering the years 1940-47, is still available. Its 713 reviews of tests and other contents are not duplicated in *The Fourth Yearbook*, and many tests reviewed in it have not been listed again in the new edition. 1047 pages.

NOTE: In 1959 the Fifth, in 1965 the Sixth, and in 1972 the Seventh Mental Measurements Yearbook (2 vols.) were published. One who can afford only one of the Yearbooks should, of course, have the latest--both volumes of the Seventh. The Sixth, Fifth, Fourth, and Third Yearbooks are still available. Even the 1938 and the 1940 Yearbooks, out of print for years, are now available again, on special order. See Catalog for

Test Service Bulletin

No. 46 THE PSYCHOLOGICAL CORPORATION January, 1954

Published from time to time in the interest of promoting greater understanding of the principles and techniques of mental measurement and its applications in guidance, personnel work, and clinical psychology, and for announcing new publications of the Test Division. Address communications to 522 Fifth Avenue, New York 36.

HAROLD G. SEASHORE, *Editor*
Director of the Test Division

ALEXANDER G. WESMAN
Associate Director of the Test Division

JEROME E. DOPPELT

MARJORIE GELINK

JAMES H. RICKS, JR.
Assistant Directors

THE CORRECTION FOR GUESSING

WHEN Pat and Mike laid down their picks and shovels and decided to apply for the job of mechanic's helper, they realized they would be competing with each other. Only one job was available. They consequently were not surprised when the personnel director asked them to take a test of mechanical comprehension to help the company decide which man would be selected.

Pat, a cautious man, carefully read the directions for the test, learned he was to choose the best answer to every question from the three choices that were given, and proceeded to take the test. He found he was quite sure of his answers to 36 of the 90 questions; for the remaining questions he could sometimes rule out one of the choices but he just could not select one answer with complete confidence. Pat felt it would be best not to try.

Mike, on the other hand, was generally more willing than Pat to take a chance. After he answered the 23 questions with which he had no difficulty, he decided to answer the remaining questions as best he could. As luck and partial information would have it, Mike managed to answer correctly 13 of the 37 items about which he had had doubts.

The results of the scoring of the test papers were 36 rights and no wrongs for Pat; 36 rights and 24 wrongs for Mike.

The test-maker had realized that people will react differently when faced with multiple-choice test questions which they cannot answer with confidence. Some will not respond to such questions; others will risk answering them. Consequently the test score was defined as the number of correct answers *minus one-half the number of wrong responses.* Thus Pat's score was 36 and Mike's score was 24.

In this instance, the correction for guessing resulted in a higher score for Pat than for Mike. Since we know how the two men took the test, it seems entirely fair for Pat to receive the higher rating. But how often do we know what has gone on in the minds of the examinees?

All people who have scored or used multiple-choice tests know that there exist several "formulas" for obtaining scores. We find among objective and semi-objective tests such different scoring formulas as the number of right answers, the number of right answers minus the number of wrong responses, Rights minus ½ Wrongs, Rights minus ⅓ Wrongs, Rights minus ¼ Wrongs, and the like. Psychometricians can usually tell, after a quick glance at the test content, what the scoring formula will be. If the test is of the completion type, the formula is the number of right answers; if the test is of the multiple-choice type, the number of right answers is often reduced by an amount equal to the number of wrong responses divided by one less than the number of options per item.

The scoring formula for a particular test is determined by applying the laws of chance in an attempt to correct for the effect of guessing on the part of the examinee. In a test made up of five-choice items, for example, the examinee may be expected to guess correctly the answer to one out of every five items he doesn't know or can't solve. The total number of right answers therefore includes the number of items answered correctly on the basis of information plus the number of correct guesses. But how does one know how many answers are guesses? To determine the number of correct guesses we make use of the number of wrong responses. Thus, for every four wrong responses it is assumed the examinee made one correct guess. Consequently, the number of correct guesses is estimated for a five-choice test by dividing the number of wrong responses by four. This is the correction for guessing and it is subtracted from the number of right answers.* Note that the basic assumption is that all wrong responses plus some of the right ones are classified as chance responses or guesses.

Usually a "guess" is interpreted as a positive statement or action based on chance. An omission or the withholding of a response is ordinarily not considered a guess. It is interesting, therefore, to find that in any given group, how much difference the so-called correction for guessing makes depends on the number of omissions rather than on the number of actual guesses (which we never know). If everyone in a group taking a test answers *all* the items, the uncorrected scores (the number of right answers) will be perfectly correlated with corrected scores which take into account the number of wrong responses. The numerical values of the corrected and uncorrected scores will of course be different but the relative positions or ranks of individuals in the group will be exactly the same. This can be demonstrated mathematically and will be true *regardless of whether the wrong answers are due to chance responses, misinformation or partially correct information.* The same situation exists if all students have the same number of omitted items, even though the specific omitted items differ from student to student. It is only when the number of omissions ranges from very few to very many that a correction factor assumes significance.

What does the "corrected" score mean? Is it an indication of the number of items to which the examinee definitely knows the answers because the number of correct guesses has been subtracted? After some consideration, one can see that the corrected score does not actually mean this. For the correction formula to be strictly applicable, the examinee must have made pure chance responses to all the items which he marked incorrectly and to some of the items which he marked correctly. For that to occur, all of the options for an item to which the examinee responds by chance must seem to him equally likely to be right. Ordinarily, if the examinee is even half awake when he is taking the test, all the options will not be equally attractive. He can probably rule out some of the options quite readily. It is also obvious that influences other than chance enter into the picture. It is entirely possible that an examinee answers an item incorrectly because he has definite misinformation on the topic or because he has partial information which misleads him. In such cases, he did not really guess at the answer, in a chance sense. Since the examinee rarely chooses purely by chance among the possible answers presented to him, the basic assumption underlying the correction for guessing is violated. In some instances the correction for guessing may overcorrect and in other instances, it may undercorrect. In general, the correction for guessing probably yields a reasonable approximation of the true situation not because of the inherent soundness of its assumptions but rather because it tends to be a compromise between too much correction and not enough correction.

If the correction for guessing is based on conditions which are practically never met, some terms and concepts regarding the meaning of test scores should be altered. For example, it is not uncommon to hear that a particular student or applicant got no more than a "chance score" on a test if he answered twenty items correctly out of a total of 100 five-choice items. It is felt that such a score is no more than the effect of chance, and the correction for guessing if he tried every item will reduce this score to zero. It is no more accurate to say that this student got a "chance score" than it is to say that a pair of loaded dice respond to the laws of chance. There is probably no such thing as a chance score on a test appropriate to the person and the situation unless the examinee is blindfolded when he takes the test. Zero scores and negative

* The formula is $R - \frac{W}{N-1}$ where R is the number of right answers, W is the number of wrong responses and N is the number of choices per item. It is sometimes remarked that the larger the number of possible answers to a question, the smaller the importance of the correction formula. On a two-choice or true-false item, chance answers will be right fifty per cent of the time; for a five-choice item, the probability is only twenty per cent; and for a sixteen-choice, such as appears in the *DAT Verbal Reasoning Test*, one can quite safely ignore the role of chance.

scores, which sometimes result from a correction for guessing, are not indications of no knowledge whatsoever regarding the materials in the test. Such scores are probably obtained through the interaction of (a) positive correct information on some items, (b) guessing and partial information and positive misinformation on other items, and (c) overcorrection for guessing.

The correction for guessing is widely used in scoring power tests and there are some situations in which its use is advisable. Some students are bold, and answer questions when they are not sure of the answers while their more timid colleagues would rather omit those questions. If the test score is simply the number of correct responses, there will be a premium for boldness. In such instances, it seems reasonable to correct the scores by subtracting a proportion of the number of wrong responses. It would, however, be more logical to call the correction factor a penalty for wrong responses than to call it a correction for guessing.

When an item is omitted in an untimed test we can generally assume that the examinee had the opportunity to read the question but, for one reason or another, refused to respond. Speed tests present a somewhat different problem. True speed tests are made up of questions which are extremely easy and the examinee will almost always answer correctly if he has the opportunity to read the item. Most of the omissions in a pure speed test are due to the fact that the examinee never got a chance to answer the items because the time was up before he could reach them. In tests of this type we usually find very few or no omissions and relatively few wrong answers between the first item and the last item attempted. Consequently there is no need for using a correction scheme. The number of right answers is entirely adequate as a score.

Many tests may best be described as a mixture of power and speed. In such tests speed is an important factor, but the items vary in difficulty and are generally arranged in order of difficulty. Between the first and the last items attempted the examinee is very likely to encounter questions which he cannot answer with certainty and he must then decide whether or not he will risk a guess. There may be considerable variation in the number of omissions up to the last item attempted and in the number of wrong responses. If this is found to be the case, the situation is similar to that found in power tests. A corrected score may then be advisable. There are, however, great differences among tests of the mixed power and speed type in the extent to which items are omitted or answered incorrectly. The actual effectiveness of a correction applied to the number of right answers must be evaluated for each test separately.

Before leaving the topic of speeded tests we should note a situation which occasionally arises. Many speeded tests are scored by merely recording the number of right answers. A test-wise examinee who knows how the test is scored may answer the questions to the best of his ability until shortly before the time is up. He may then hastily record an answer to each of the remaining items without even stopping to read the questions. He is thus almost bound to pick up some points of score without any danger of incurring a penalty. If this kind of test-taking occurs with any frequency, it would be advisable to apply a correction to the scores.

It is the fond hope of those who construct power tests that the examinees will answer all the items. If this were to happen, there would be no need for correcting scores. We know, however, that there are differences among examinees in their willingness to leave items unanswered. It is likely that more people might be induced to answer more items if the directions for the test stated that omissions would be counted as wrong responses or if all were encouraged to guess whenever they did not know the answers. Such directions would doubtless disturb those educators who feel that encouraging students to guess makes for loose thinking and disrespectful attitudes toward learning. This view may have validity if tests are being used for moralistic or character-building purposes alone, but this is rare. Usually a test's essential function is that of a measuring instrument and as such it should be kept as uncontaminated as possible. One source of contamination is the matter of boldness vs. caution in taking the test. The imposition of a penalty for wrong answers is an attempt to control this type of contamination. More effective control would be achieved if all students were encouraged to be equally bold, so to speak, by answering every item.

What can be said, in summary, about the correction for guessing? In the first place, "correction for guessing" is essentially a misnomer; the correction could more properly be called a penalty for answering wrong. Some of these wrong answers may have been given in accordance with the laws of chance but more of them probably are based on misinformation or partial information.

Second, the basic assumption underlying the correction for guessing is the concept of the "chance score." Thus one expects a proportion of the number of items to be answered correctly on the basis of chance. This concept is misleading and may make it appear that the examinee knows the answers to fewer questions than he really does know.

Third, the correction for guessing makes a difference in the relative positions of individuals in a group if there is considerable variation in the number of omitted items. To eliminate a premium for willingness to answer items, it seems advisable to use corrected scores. It should be remembered, however, that such corrected scores are an attempt to rule out the effects of differential boldness in taking the test rather than a method for getting a true picture of the examinee's knowledge.

Fourth, when a comparison of the corrected and uncorrected scores on a power test shows considerable discrepancies in relative standing of the examinees, the question is not which type of score should be used. The question is whether or not the test is really a power test and whether it is appropriate for use with the group.

The fundamental purpose in giving a test is to obtain samples of behavior which will permit comparisons with respect to some reasonably well defined attribute among individuals in the group tested. Effective discrimination among the examinees must be demonstrated for each test in specific situations. There is no reason to believe that any scoring formula contributes materially to a test's discriminating power.

Many published tests require use of a correction formula to obtain the score. The user of such tests must necessarily abide by the scoring instructions since otherwise he cannot compare his scores with the norms. In making his own objective tests, the teacher or personnel man need not feel that a correction for guessing is essential to the construction of a good test. Reliability and validity may still be obtained with either corrected or uncorrected scores.—J.E.D.